DOCUMENT RESUME

ED 174 639 TM 009 247

AUTHOR Meredith, Colin
TITLE A Monte Carlo Investigation of Several Procedures for Testing the Significance of Individual Discriminant Functions.
PUB DATE Apr 79
NOTE 50p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Hypothesis Testing; Simulation; *Statistical Analysis; Tests of Significance
IDENTIFIERS Monte Carlo Method

ABSTRACT

The problem of determing how many significant discriminant functions are present in a given data set for a one-way, fixed-effects multivariate analysis of variance design is studied using a Monte Carlo procedure. A variety of procedures, including the popular partitioned-U procedure, are compared with respect to their Type I error rates and power to detect true population differences. Results of the study lead to the recommendation of the cautious use of the partitioned-U procedure. (JKS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174639

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A MONTE CARLO INVESTIGATION OF SEVERAL PROCEDURES FOR TESTING THE SIGNIFICANCE OF INDIVIDUAL DISCRIMINANT FUNCTIONS

COLIN MEREDITH
STEVENSON AND KELLOGG

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Colin Meredith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM"

Paper presented at the 1979 Annual Meeting of the American Educational Research Association, San Francisco, California, April 8-12, 1979.

SESSION # 24.20

TM009 247

This paper is concerned with the problem of determining, for a one-way, fixed effects multivariate analysis of variance (MANOVA) design, how many significant discriminant functions are present in a data set. The most frequently recommended procedure (e.g., Tatsuoka, 1971; Overall and Klett, 1972; Timm, 1975) for determining how many of a possible s discriminant functions are significant (where s is equal to the lesser of p, the number of dependent variables, and df_h, the hypothesis degrees of freedom) is the partitioned-U procedure, first proposed by Bartlett (1941). Despite its seeming popularity and acceptance, the partitioned-U procedure is not without its critics. Notably, Harris (1975, 1976) contends that the procedure is both logically and statistically invalid. Specifically, Harris (1976) claims that the procedure neither tests the first root, nor takes into account the possibility that the size order of the sample roots may not correspond with the order of the population roots. The purpose of the research reported here is to compare a number of procedures purported to test the hypotheses alluded to above, in terms of their type I error rates and power to detect true population differences.

The past decade has seen a substantial growth in the application of multivariate analysis procedures by social scientists. While the work of researchers such as Olson (1974, 1976) has provided guidelines for choosing a 'best test' from among a number of alternative overall MANOVA test statistics, such guidelines are frequently not available to the analyst uncertain about which procedure to apply after the overall MANOVA null hypothesis has been rejected. A number of options are available

including univariate analysis of variance on each of the p dependent variables (Cramer and Bock, 1966), calculation of step-down or residual F-statistics (Roy and Bargmann, 1958), and discriminant function analysis (for a review of this procedure, see Huberty, 1975). It is the application of this third procedure, discriminant function analysis, as a follow-up procedure to MANOVA, which is investigated by the research reported here. A recent general review of these procedures is provided by Miller (1977).

The model under consideration is the one-way fixed effects design in which p (>1) normally distributed response measures are available for n experimental units in each of q treatment groups. The overall MANOVA null hypothesis for such a design can be stated as:

$$H_{ov} : \mu_1 = \mu_2 = \cdots = \mu_q ,$$

where the μ_q are the p-variate mean parameter vectors associated with each of the q levels of the treatment. Importantly, all of the frequently-applied MANOVA test criteria are functions of the latent roots, or eigenvalues of the matrix product $E^{-1}H$ where H and E are pxp sums of squares and cross products (SSCP) matrices defined in the one-way case as

$$H = \sum_{i=1}^{q} n_i \, (\bar{Y}_i - \bar{Y}) \, (\bar{Y}_i - \bar{Y})' \qquad (1.1)$$

$$E = \sum_{i=1}^{q} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)(Y_{ij} - \bar{Y})' \qquad (1.2)$$

where Y_{ij} is the jth of n_i observation vectors in group i, $\bar{Y}_i$ is the mean vector for the ith group and $\bar{Y}$ is the grand mean vector. Examples of

overall MANOVA test criteria which are functions of the eigenvalues of $E^{-1}H$ are Roy's largest root criterion (Roy, 1945), Wilks' likelihood ratio (Wilks, 1932), the Lawley-Hotelling trace criterion (Lawley, 1938; Hotelling, 1951) and Pillai's trace criterion (Pillai, 1955). The distribution theory of the eigenvalues used in these tests is reviewed and discussed in some detail by Pillai (1976, 1977). For a comparison of these procedures in terms of their robustness to violation of certain of their distributional assumptions, the reader is referred to Olson (1974).

As is clear from the statement of the overall MANOVA null hypothesis above, a significant overall statistic implies only that, in terms of the parameter mean vectors, some group (or groups) differ from others on some variate (or variates). No indication is given as to where exactly the population differences are to be found. Discriminant analysis addresses this question by determining the number of distinct ways in which the q groups differ significantly on the p variates, sometimes referred to as dimensions of difference (Huberty, 1975). This is accomplished by forming linear composites of the original variables, called discriminant functions, which weight the variables in a manner which maximizes the ratio of between-group to within-group variation while taking into account the intercorrelations among the dependent variables. As many as s functions are possible, each of which is associated with one of the latent roots of $E^{-1}H$. The maximization procedure employed results in the extraction of the sample roots (denoted ℓ_1) in decreasing order of magnitude, with successively derived functions uncorrelated (in the total sample) with those already derived. The problem then becomes one of determining how many of these sample roots are statistically significant, meaning that they represent reliable dimensions of difference in the population. Huberty (1975), in his comprehensive review of discriminant analysis, comments that "in the literature, the

process of testing the significance of a function has been lacking in clarity" (p.548). To date, a number of procedures have been proposed in response to this situation. These considered here are:

1) The partitioned-U procedure due to Bartlett (1941);

2) A variation on Roy's largest root test proposed by Finn (1974);

3) Another variation on Roy's test proposed by Harris (1976);

4) A variation of Bartlett's procedure proposed by Marriott (1952); and

5) A combination of Bartlett's and Roy's tests proposed by Harris (1976).

Before discussing the details of these procedures, it will be convenient to introduce the following notation. Let the pxp matrix M, the parametric counterpart of H, be defined as

$$M = \sum_{i=1}^{q} n_i (\mu_i - \mu)(\mu_i - \mu)' \qquad (1.3)$$

where μ_i is the population mean vector for the ith of q treatments, μ is the grand mean vector in the population, and n_i is the number of observations in the ith group. Further, let Σ denote the pxp population SSCP matrix common to each of the q levels of the treatment populations, estimated by E, and let Ξ denote the pxp population variance-covariance matrix. It is the nonzero latent roots of $\Sigma^{-1}M$, denoted $\lambda_1, \lambda_2, \ldots, \lambda_r$, (where r is the rank of $\Sigma^{-1}M$) which are estimated by the roots of $E^{-1}H$ ($\ell_1, \ell_2, \ldots, \ell_r$), and which correspond to the true population discriminant functions. The following sections describe in some detail the five procedures mentioned above, and present statements of the null hypotheses tested by each in terms of the population roots.

The Partitioned-U Procedure. One approach to testing the significance of a given value of U (the Wilks' lambda (Wilks, 1932) overall MANOVA statistic), where U is defined as

$$U = \prod_{i=1}^{s} \frac{1}{1 + \ell_i} \tag{1.4}$$

involves calculating Bartlett's V statistic as

$$V = [N - 1 - (p+q) / 2] \sum_{i=1}^{s} \ln (1+\ell_i) \tag{1.5}$$

where N is the total sample size, and p and q are as defined previously. V is distributed approximately as a chi-square with p(q-1) degrees of freedom when used as a test of the overall MANOVA null hypothesis. Since the successive terms $\ln(1+\ell_i)$ in expression (1.5) are statistically independent (assuming multivariate normality of the original p variables) the additive components of V are each approximately distributed as chi-square variates with p+q-2i degrees of freedom where i is the number of components removed from the expression for V.

The partitioned-U procedure per se involves the following steps (as outlined by Harris, 1976):

(1) Test the overall null hypothesis by comparing V to χ^2 with p(q-1) degrees of freedom. If H_o overall is not rejected, it is concluded that none of the discriminant functions are statistically significant, and testing stops. If, however, this null hypothesis is rejected, then at least the first discriminant function is considered significant, and the researcher proceeds to step 2.

(2) Test the significance of the residual after successively removing the chi-square components (i.e., $\ln(1+\ell_i)$) corresponding to the first

i discriminant functions. As soon as the residual becomes smaller than the critical value of the appropriate chi-square distribution, it is concluded that only these first i functions are significant (at that α level). The successive test statistics and their associated degrees of freedom (as shown in Tatsuoka, 1971, p. 165) may be summarized as follows:

Residual after Removing	Approximate χ^2 - Statistic	Degrees of Freedom
First discriminant function	$V - V_1$	$p(q-1) - (p+q-2)$ $= (p-1)(q-2)$
First 2 discriminant function	$V - V_1 - V_2$	$(p-1)(q-2) - (p+q-4)$ $= (p-2)(q-3)$
First 3 discriminant function	$V - V_1 - V_2 - V_3$	$(p-2)(q-3) - (p+q-6)$ $= (p-3)(q-4)$
.	.	.
.	.	.

The null hypothesis tested by this procedure is stated succinctly by Kshirsagar (1972) as follows: for any value of i greater than zero (meaning <u>not</u> the overall test) the first i population roots are nonzero, while the remaining s-i roots are equal to zero. Hence, this procedure implicitly rejects the hypothesis that $\lambda_1 = 0$ when the overall statistic is found to be significant.

Finn's (1974) variation of Roy's criterion. As mentioned above, Roy's largest root criterion defined in terms of the roots of $E^{-1}H$ as

$$R = \frac{\ell_1}{\ell_1 + 1} \tag{1.6}$$

was developed as a test of the overall MANOVA null hypothesis. This fact notwithstanding, Finn (1974) and Harris (see below) have recommended

using this statistic to test null hypotheses of the form $H_o : \lambda_i = 0$, when the previous (larger) i-1 sample roots have been tested. Finn's (1974) procedure involves reducing two of the components of the degrees of freedom by one for each prior root tested. As a test of the overall hypothesis, R is compared to the critical value of the greatest characteristic root distribution (Pillai, 1965) with the following parameters:

$$s = \min (p, q-1) \tag{1.7}$$

$$m = (\left| p-df_h \right| - 1) / 2 \tag{1.8}$$

$$n = (df_e - p - 1) / 2 \tag{1.9}$$

where df_h and df_e are the degrees of freedom for hypothesis and error, respectively. As a test of the significance of discriminant functions other than the first (presumably proven significant by the overall test) Finn would calculate R as

$$R = \frac{\ell_i}{\ell_i + 1} \tag{1.10}$$

and evaluate this statistic against the greatest root distribution with df_h and df_e in the m and n arguments, respectively, reduced by one for each prior root tested. Testing would terminate with the finding of a nonsignificant statistic. The null hypothesis tested at any given stage is $H_o : \lambda_i = 0$, given that the previous i-1 roots have been found to be nonzero.

Harris' (1976) variation of Roy's criterion. This procedure is similar in all respects to Finn's (1974) procedure except that, rather than reducing df_e and df_h, Harris recommends reducing s by one for each prior

root tested, while leaving the m and n arguments unchanged.

Marriott's (1952) Chi-square Procedure. Marriott (1952) proposed a variation on the partitioned-U procedure which both takes into account the possibility that the ordering of the sample roots does not correspond to the ordering of the population roots, and provides a direct test of λ_1 . The procedure involves the following steps:

1) Test the hypothesis $H_o : \lambda_1 = 0$ by evaluating the significance of

$$V_1 = - [N - 1 - (p+q) / 2] \ln \ell_1 \qquad (1.11)$$

as a χ^2 variate with degrees of freedom equal to

$$p + q - 2 + 1/2 [(p-1) (q-2)]^{2/3} \qquad (1.12)$$

2) Test the hypothesis that $\lambda_{i+1} = 0$ when the previous i roots have been shown to be significant by evaluating

$$V_i = - [N - 1 - (p-i + q-i) / 2] \ln \ell_i \qquad (1.13)$$

as a χ^2 variate with degrees of freedom

$$p - i + q - i - 2 + 1/2 [(p-i-1) (q-i-2)]^{2/3} \qquad (1.14)$$

Step 2 is repeated with i increasing until either nonsignificance is obtained or i=s. Marriott (1952) reports some data indicating that this procedure performs well when p is small (4 or less) but may become liberal as p increases. The validity of these assertions has not, before the present study, been put to an empirical test.

Harris' (1976) Mixed Procedure. As an alternative to the use of Roy's criterion for testing roots other than the first (see above) Harris (1976) has proposed a procedure which combines the partitioned-U test and a variation on the largest root test. The steps in this 'mixed' procedure are as follows:

1) Test the overall null hypothesis using Bartlett's V with p(q-1) degrees of freedom. If this null hypothesis is rejected, conclude that there is at least one significant discriminant function, and proceed to step (2). If the observed V does not exceed the critical value, testing stops.

2) Test the significance of the largest population root by comparing R_1 to the tabled critical value of the greatest characteristic root distribution with values for s, m and n calculated in the manner described above. If this test achieves significance, conclude that the population value of λ_1 does not equal zero, and proceed to step (3). Otherwise conclude that the nature of the relationship between group membership and the dependent variables cannot be specified, and stop testing.

3) Test the hypothesis that the remaining s-i population roots are all equal to zero by comparing

$$V = [N - 1 - (p+q) / 2] \sum_{K=i+1}^{s} \ln (1+\ell_K) \qquad (1.15)$$

to χ^2_α with (p-i)(q-i-1) degrees of freedom. If this test fails to achieve significance, conclude that only the first i population roots are nonzero. If it achieves significance, conclude that not all of the last s-i population roots are zero, and proceed to step (4).

4) Test the null hypothesis that the i + 1th population root is equal to zero by comparing R_{i+1} to the greatest characteristic root distribution with s' = s-i. If nonsignificant conclude that the nature of the relationship associated with the last s-i roots cannot be specified. If significant, conclude that the i + 1th discriminant function is nonzero (in the population), and return to step (3). Steps (3) and (4) are repeated until a nonsignificant test terminates the procedure, or until i=s.

To date, little, if any, empirical investigation of these procedures has been reported in the literature. Bartlett's (1941) discussion of the partitioned-U procedure is largely analytical in content. Harris (1976) comments that his proposed variation of Roy's test may be conservative in its evaluation of roots other than the first, while the mixed procedure may prove frustrating to use due to the fact that different sampling distributions are referenced at different stages of the procedure, resulting in potentially contradictory conclusions. Marriott (1952) investigated the behaviour of his proposed procedure for selected levels of p and q, and noted some overestimation of significance levels as p increased. These comments pertain solely to the type I error rate levels of these tests. No individual or comparative data are at present available regarding the power of any of these tests to detect the existence of nonzero population roots given a set of sample data. It will be convenient at this point to introduce the MANOVA noncentrality parameter, a standardized measure of the distance between group centroids in the population. Given a

pxp parameter hypothesis SSCP matrix, M as defined above, and the population covariance matrix, Ξ , let

$$H^* = M \Xi^{-1} . \qquad (1.16)$$

The noncentrality parameter is the sum of the eigenvalues, h_j (for j = 1,2,...,p) of H*, or simply tr(H*), the trace of H* (Schatzoff, 1966b; Olson, 1973).

In response to the need for both of the types of information mentioned above, the research reported here comprised two distinct phases. In study I, the procedures were compared in terms of their type I error rates for various combinations of p, q and n. In study II, varying degrees and patterns of multivariate noncentrality were incorporated in the simulation procedure, and the procedures were compared in terms of their power to detect these population discriminant functions. The simulation procedure did not incorporate violations of the distributional assumptions of the overall tests.

Method

Even in the absence of assumption violation, the researcher who wishes to investigate almost any aspect of multivariate analysis is faced with the problem of choosing from among a seemingly infinite number of combinations of independent variables. Central to this problem in the present context of a one-way MANOVA is the fact that, subject to the constraint of s always being limited to the lesser of p and df_h, any given set of population roots can result from an infinite number of combinations of p, q, n, that form the q p-variate population mean vectors, and the

common population covariance matrix. Even holding all but the mean vectors constant, there exists an infinite number of sets of mean vectors which, in combination with the other factors, will yield a given set of roots. This problem makes simulation of a particular set of roots by starting with the raw data a somewhat confusing and arduous task (although not an impossible task - see Olson, 1973). In order to circumvent this problem, the data generation procedure utilized by Schatzoff (1966b) was employed in the present study. Details of this procedure are presented below, following an enumeration and description of the independent variables investigated.

1. Number of dependent variables: p. The levels investigated were 2, 3, 6 and 10 in the null situations (study I) and 2, 3 and 6 in the noncentral cases (study II).

2. Number of groups: q. All of the simulated experiments had a one-way design with 3, 6 or 10 levels of the treatment factor.

3. Number of observations per group: n. The levels of this variable examined were 5, 10 and 50 with all groups in any simulated experiment being equal in size.

The following factors relate to the manipulation of multivariate noncentrality and were, accordingly, investigated only in study II.

4. Number of nonzero population roots: r. Subject to the limitation that r not exceed s for any combination of p and q, experiments were simulated in which r, the population dimensionality, was equal to 1, 2, 3 and 5.

5. Degree of multivariate noncentrality: $tr(H^*)$. The values of this parameter investigated were 0 (indicating the null case, as in study

1), 10, 40 and 90.

6. Pattern of noncentrality. The following patterns of the roots of H* were investigated:

a) concentrated: in which only the first root of H* is nonzero and is therefore itself equal to tr(H*).

b) equal: in which the r nonzero roots of H* are equal in size and sum to tr(H*).

c) linear: in which the size of the roots from 1 to r decreases by a constant amount in a linear fashion.

d) geometric: in which the size of the roots from 1 to r decreases by half in a geometric fashion.

The values of the factor levels were chosen to facilitate comparison of the results of the present study with those of Olson (1973, 1976) and Schatzoff (1966b). Table 1 displays the levels of these factors investigated in each of the three studies in question, as well as in the recent study reported by Mendoza, Markos and Gonter (1978).

The various combinations of these factors examined in the present investigation allowed for 36 (pxqxn) distant cases to be simulated in study I, while 405 distinct factor combinations were possible in the noncentral situation. In each of the null cases, except those in which n=50, for p at least 6, or with p=q=n=10, 5000 replications were generated. In all remaining cases 1000 replications were run.

Table 1

The values of the independent variables investigated by Olson (1973), Schatzoff (1966b) Meredith (1978) and Mendoza et al (1978)

Factor	Study			
	Olson (1973)	Schatzoff (1966b)	Meredith (1978)	Mendoza et at (1978)
p	2,3,6,10	2,4,6	2,3,6,10	6
q	2,3,6,10	3,5,7	3,6,10	7
n	5,10,50	25,100*	5,10,50	25,50,100,200,500**
tr(H*)	0,10,40,90	12,24	0,10,40,90	0,.79,1.01,2.87,3.76, 15.33
Pattern	concentrated	equal	concentrated	three different nonzero variates in each condition
	diffuse***	linear geometric	equal geometric	

* error degrees of freedom
** total sample size
*** differences on all canonical variates

The data generation procedure employed was that outlined by Schatzoff (1966b). Specifically, a p x df_e matrix, T_1, is generated the elements of which are independently and identically distributed N (0,1) random variables. The sample error SSCP matrix E is then computed as

$$E = T_1 T_1'$$

In similar fashion, a p x df_h matrix, T_2, is generated, also composed of N (0,1) variates, and the hypothesis SSCP matrix, H, is computed as

$$H = T_2 T_2' .$$

This definition of H applies to the null case only. In the noncentral case values equal to the square roots of h_j, the eigenvalues of H*, are added to the elements of the principal diagonal of T_2 prior to computing H, in order to achieve the desired degree and form of multivariate noncentrality. The theory underlying this procedure is discussed by Dempster (1969). Briefly, it relies upon the fact that if H and E are independent random pxp matrices distributed as Wishart matrices with q-1 and N-q degrees of freedom respectively, then the quantity

$$\frac{|E|}{|H + E|}$$

has the desired distribution of Wilks' Λ statistic (Dempster, 1969, p. 311), and the roots of $E^{-1}H$ are the desired estimates of $\lambda_1, \lambda_2, \ldots \lambda_s$ as defined above.

Once $E^{-1}H$ was computed for each replication, the sample eigenvalues were obtained as the solutions to the determinantal equation

$$\left| E^{-1}H - \ell I \right| = 0$$

Values of the five statistics under consideration as well as Bartlett's V and Roy's overall MANOVA tests were calculated as described above. Counts were kept of the number of experiments in which the statistics exceeded the tabled critical values. Successively smaller ℓ_i were tested either until a nonsignificant result was obtained or until s roots had been evaluated. These procedures were carried out for nominal significance (α) levels of .05 and .01. The raw data were converted to percentage

exceedance data in which the rejection count for the tests of any given root are expressed as a percentage of the number of samples in which the test of next largest root was found significant. For example, if a given procedure rejected $H_o : \lambda_3 = 0$ in 500 samples, and then rejected $H_o : \lambda_4 = 0$ in 200 of these samples, then the percentage exceedance for this procedure as a test of the fourth population root would be (200/500) x 10 = 40%.

In all, 575,000 simulated experiments were generated and analyzed in the present study.

Results

Certain data relevant to the practicability of three of the procedures become evident in advance of any simulation. With respect to Harris' (1976) mixed procedure, and Harris' (1976) variation of Roy's test (hereafter these procedures will be referred to as MX and H, respectively), the procedures could not be applied exactly as originally described since s will necessarily equal 1 for the test of the last (sth) root, and no tables of the greatest characteristic root are available for s=1. This is not surprising, in that any situation where s=1 is analyzable as a multiple regression problem. In the present study, MX and H tested the significance of the sth root as an F-statistic with k and N-k-1 degrees of freedom, where N is the total sample size, and k is the absolute value of the difference between p and $df_h + 1$. This procedure is outlined by Marriott (1974) and falls within the general format for testing the significance of multiple correlation coefficients.

A second pre-simulation finding concerned Finn's (1974) variation

of Roy's test (hereafter referred to as F). It was discovered in the process of obtaining the tabled critical values, that reducing df_h in the m argument could result in the critical value increasing in some cases across roots, decreasing in other cases, and even decreasing, and then increasing again across certain sets of roots. The size pattern of m across roots (and thence, of the critical values) depended upon the relative sizes of p and q. For example, if p=3 and q=6, the value of the m parameter for the test of the first root is 1/2, for the second root m=0, and for the third root m = - 1/2. Here, as would be expected, the size of the critical value decreases as successively smaller roots are tested. If, however, p=6 and q=3, then m_1 = 1 1/2 while m_2 = 2. Finally, if p=6 and q=10, then the values of m associated with the critical values used to test the successively smaller sample roots are 1, 1/2, 0, -1/2, 0, 1/2, a nonmonotonic sequence. This serious logical inconsistency in F resulted in this procedure being eliminated from further consideration in the present study.

The principal mode of presentation of the results of studies I and II is via the use of graphs depicting the behaviour of the four procedures in various situations. As an aid to interpreting the figures, estimates of the variability inherent in the data were calculated in two different ways. In the first, a factorial analysis of variance was performed on the percentage exceedance data for tests of the first root in the null case at a nominal significance level of .05. The factors in the analysis were: p, q, n and the test. The four-way interaction, used as a residual error mean square, was found to be 0.128 based on 36

degrees of freedom. (A summary table of this ANOVA as well as tables of raw data are available in a longer version of this report (Meredith, 1978).) The approximate standard error of a point in the graphs is then $\sqrt{.128\%} = .37\%$. Caution is therefore recommended when drawing conclusions based on differences of less than .75 percentage points. Olson (1973) points out that while such estimates are useful for assessing the effects of the various independent variables within a study, they do not address the issue of how accurately the data are estimating theoretical values. For such comparisons it is useful to calculate the standard errors of the proportional exceedance rates as $((\alpha)(1-\alpha)/NR)^{1/2}$ where NR is the number of replications, and α reflects the proportion of the replicates exceeding the critical value. Hence, when $\alpha = .05$ and 5000 replications were generated, the standard error of the percentage data is .31. In cases where 1000 replications were generated, the standard error is .69. When inspecting figures for tests of the second root in the null case, it should be remembered that the number of replications is really NR times α, since tests of the second root only follow rejections of the test of λ_1, of which approximately α x NR are expected. For example, if NR=5000 and $\alpha = .05$, then the approximate number of expected replications for the test of λ_2 is 5000 x .05 = 250, resulting in an approximate standard error for these data of $(250(.05)(.95))^{1/2} = 1.38$. Accordingly, differences between the tests in the figures for λ_2 at a nominal significance level of .05 should be interpreted cautiously if they are less than approximately three percentage points.

Study I. In this section are reported the results for the null conditions, i.e., the cases of no nonzero population eigenvalues, where the value of tr(H*), the noncentrality parameter, is equal to zero. Figures 1 through 3 present percentage exceedances of tabled critical values for the tests of the first two sample roots at a nominal significance level of .05. Dealing first with the tests of λ_1, inspection of figures 1(a), (b) and (c) reveals that, in most cases, the partitioned-U (P-U) and H were found to perform at around the 5% level, meaning that the empirical rejection rate corresponded well with the theoretical level. In contrast, Marriott's (1952) procedure (MR) and MX were found in general to be conservative, meaning that empirical α was below the theoretical level. The satisfactory performance of P-U and H was not surprising since these procedures are in fact the overall V and Roy's overall test, procedures whose operating characteristics have been well documented. The consistency of the difference between P-U and H, and MR and MX is apparent in figures 1(a), (b) and (c). To the extent that it is possible to make general statements about the individual independent variables under investigation here, the following conclusions seem warranted. The effect of increasing either of p or q is, in general, to lower the rejection rates, especially for MR and MX. The most marked effect of increasing n, the cell size, is to improve the performance of P-U especially when p is large relative to q, as in the cases of p=6, q=3 and p=10, q=3. This finding is not unexpected in that P-U was developed as a large sample procedure. The effect of significance level was entirely predictable in that rejection rates were lower for the

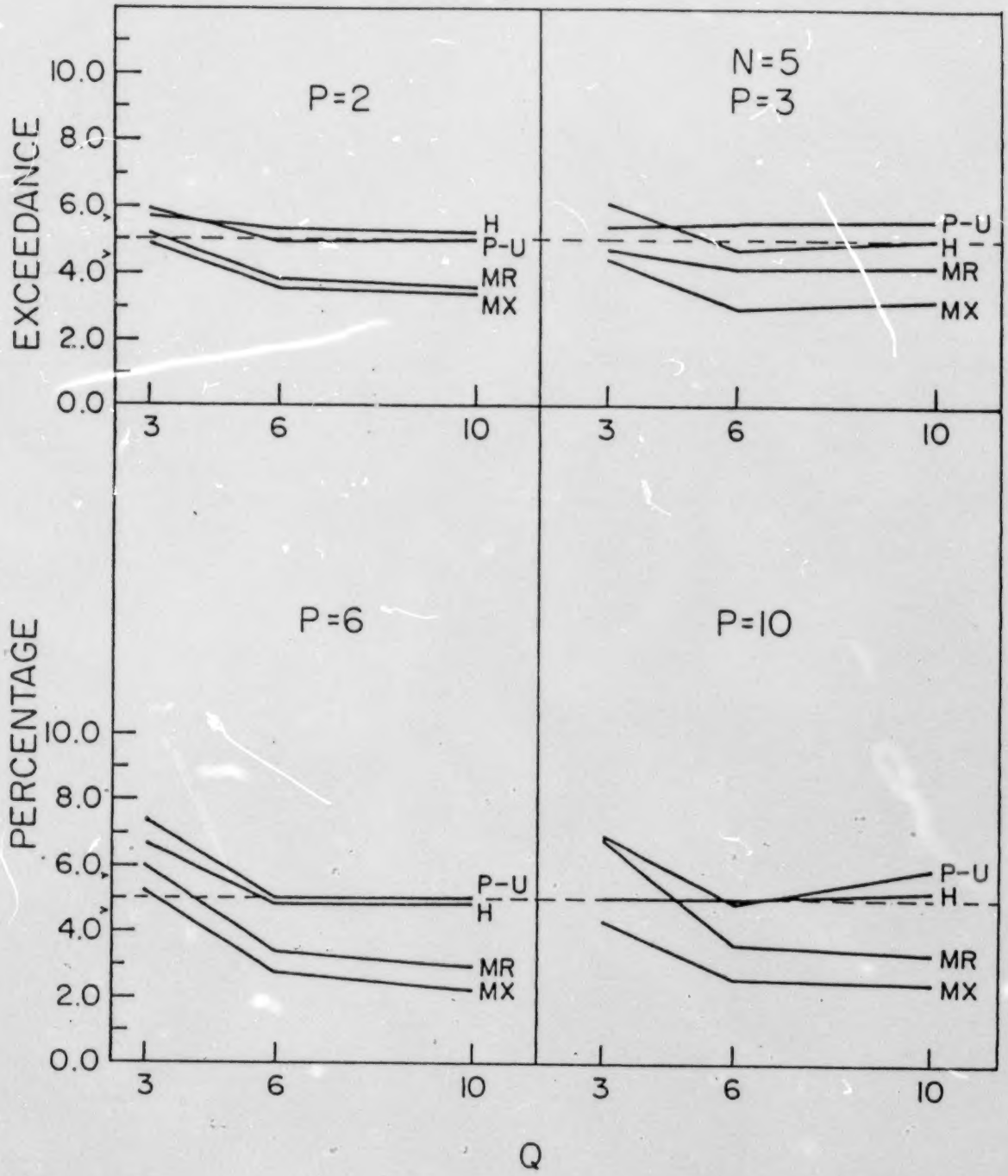

Figure 1(a). Percentage exceedance of theoretical critical values of P-U, H MX, and MR as tests of the first root in the null case for combinations of p and q given n=5. The nominal significance level of 5% is indicated by the dotted line, while the arrows beside the ordinate denote the boundaries of an approximately two standard error of proportion region around 5%.

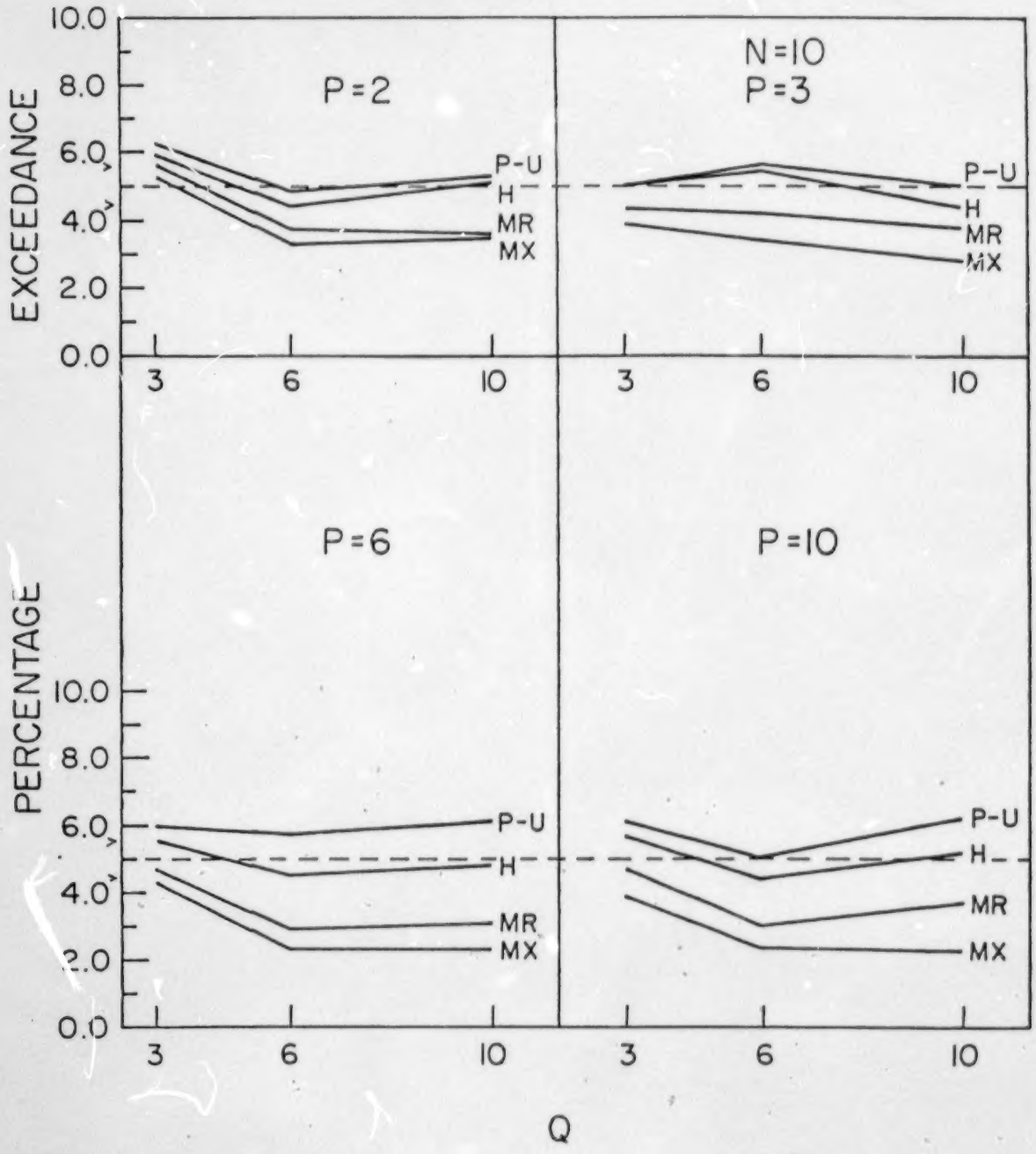

Figure 1(b). Percentage exceedance of theoretical critical values of P-U, H, MX, and MR as tests of the first root in the null case for combinations of p and q given n=10. The nominal significance level of 5% is indicated by the dotted line, while the arrows beside the ordinate denote the boundaries of an approximately two standard error of proportion region around 5%.

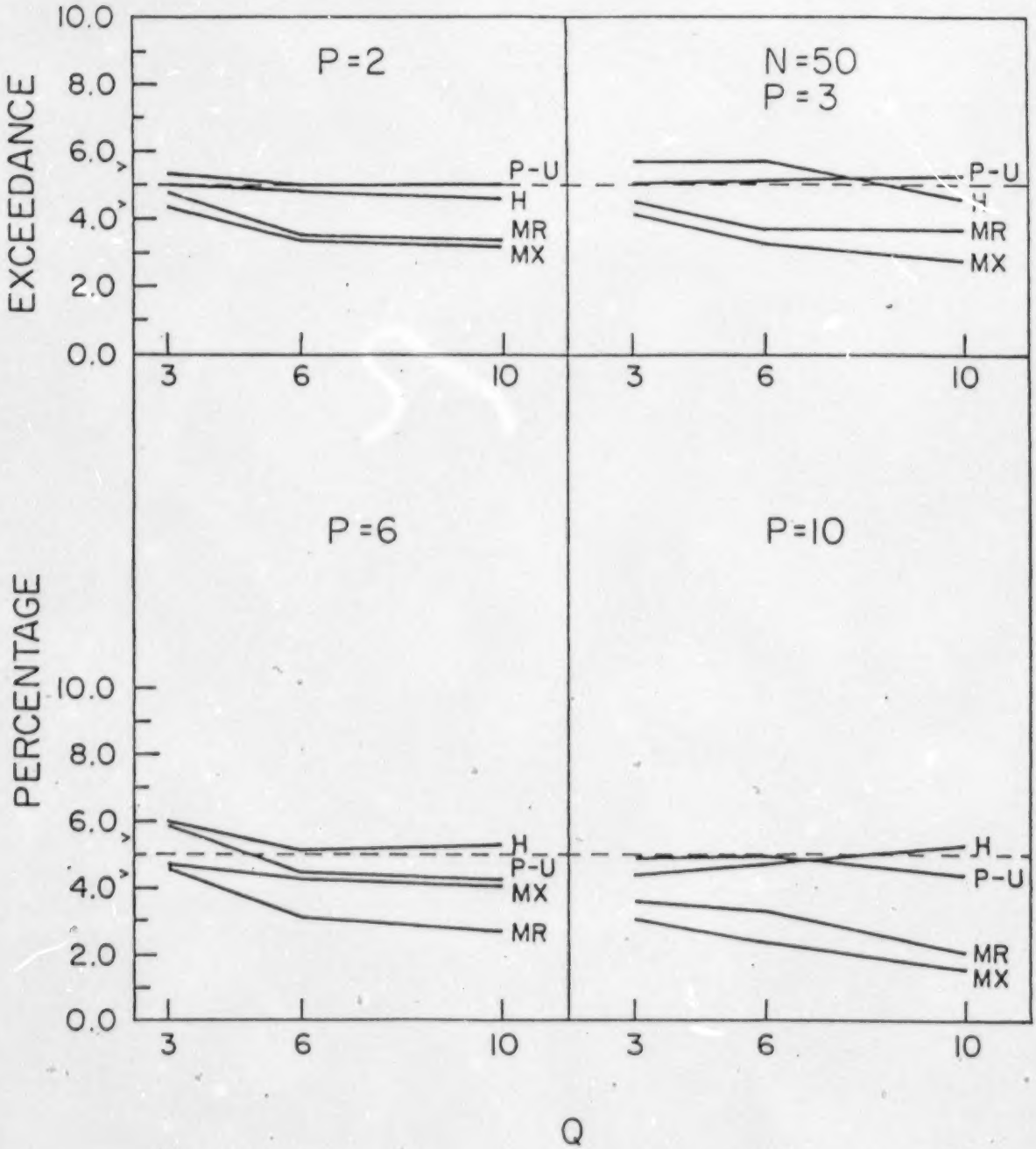

Figure 1(c). Percentage exceedance of theoretical critical values of P-U, H, MX, and MR as tests of the first root in the null case for combinations of p and q given n=50. The nominal significance level of 5% is indicated by the dotted line, while the arrows beside the ordinate denote the boundaries of an approximately two standard error of proportion region around 5%.

smaller value of α. Since the findings for $\alpha = .01$ were similar in form to those for .05, only the .05 results are presented since the rejection rates at .05 are less susceptible to floor effects.

It was thought possible that the effects of changing p and q on the tests might be due to the joint influence of these factors on s, the maximum possible number of roots. Accordingly, figure 2 was drawn to show for levels of n, the effect of s on the percentage exceedance rates of the four procedures. Inspection of figure 2 serves to confirm the general finding noted above that increasing the maximum possible dimensionality of a problem lowers the rejection rates of the procedures, especially MR and MX. Further, it was found that within values of s, the rejection rate was a decreasing function of q, the number of groups, and did not vary consistently with changes in p. This may well have been due to the influence of q on df_e.

The performance of the four procedures as tests of the second discriminant function revealed a considerable change in their relative alignment. Figures 3(a), (b), and (c) present the percentage exceedance data at a nominal significance level of .05 for the four procedures for all 36 combinations of p, q and n investigated. Whereas for λ_1, the performance of P-U and H was in sharp contrast with that of MR and MX, inspection of figure 3 reveals that, for λ_2, it is P-U and MR which differ consistently from the other two tests. Another general difference between the results for λ_2 and those for λ_1 is the lack of smoothness in the curves for λ_2 compared to those for λ_1. This is surely due at least in part to the fact that the λ_1 results were taken as proportions of 5000 (or 1000 for selected conditions) while the λ_2

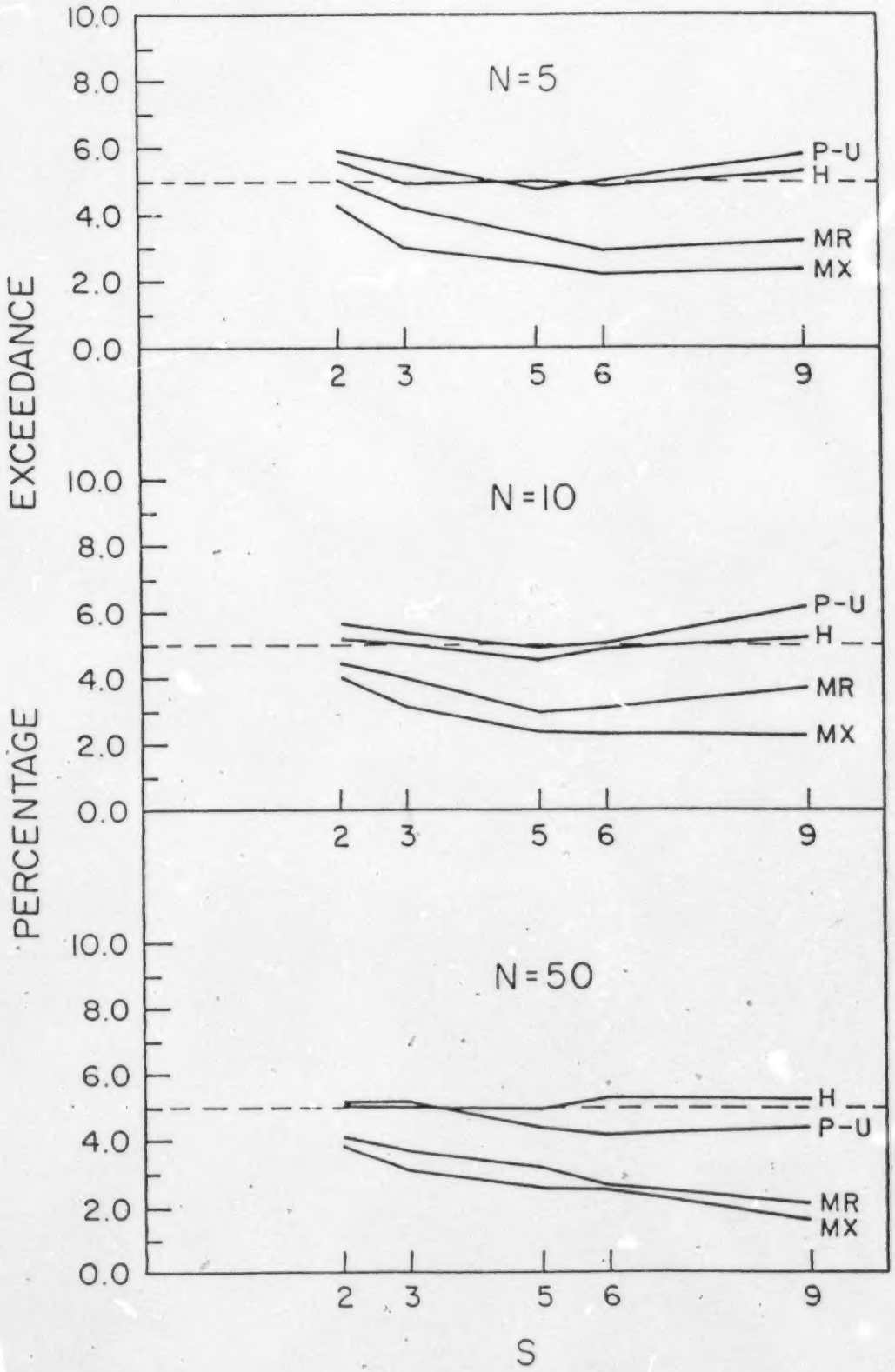

Figure 2. Percentage exceedance of theoretical critical values of P-U, H, MX, and MR as tests of the first root in the null case as a function of s for levels of n. The dotted line denotes the 5% level.

results were taken as proportions of 250 or 50 (5% of 5000 and 1000, respectively). Hence the greater variability of the estimates and concomitant lack of smoothness in the curves.

The caveat notwithstanding, the following conclusions regarding the performance of these procedures as tests of the second discriminant function seem warranted. The reader is referred to figures 3(a), (b), and (c) for visual corroboration of these conclusions. As mentioned above, P-U and MR are performing similarly as are MX and H. In general, the rejection rates of P-U and MR are closer to the nominal level of 5% than are the rates of MX and H. The rates for MX and H are generally conservative, particularly for larger values of q. In contrast to this finding is the highly anticonservative performance of P-U and MR in the condition where p=10, q=3 and n=5 (percentage exceedances of 12% and 27%, respectively), even allowing for a standard error of these points of 1.38. The liberality of P-U may be due to the fact that it is a large sample test, making it susceptible to sampling error when n is small, or when p approaches df_e, as in this case. In terms of the general effects of the independent variables individually, percentage exceedance was found to be a decreasing function of p, q, s and n.

The tests of λ_2 at $\alpha = .01$ showed very few rejections by any of the tests since, even with 5000 replications, if the nominal rate was operating one would expect only .01 x .01 x 5000 = .5 rejections of the null hypothesis for λ_2. The results for tests of roots 3 through 5 revealed only a few rejections by P-U and MR for λ_3 at $\alpha = .05$. No other rejections were detected.

Although these data address the issue of the relative propensity

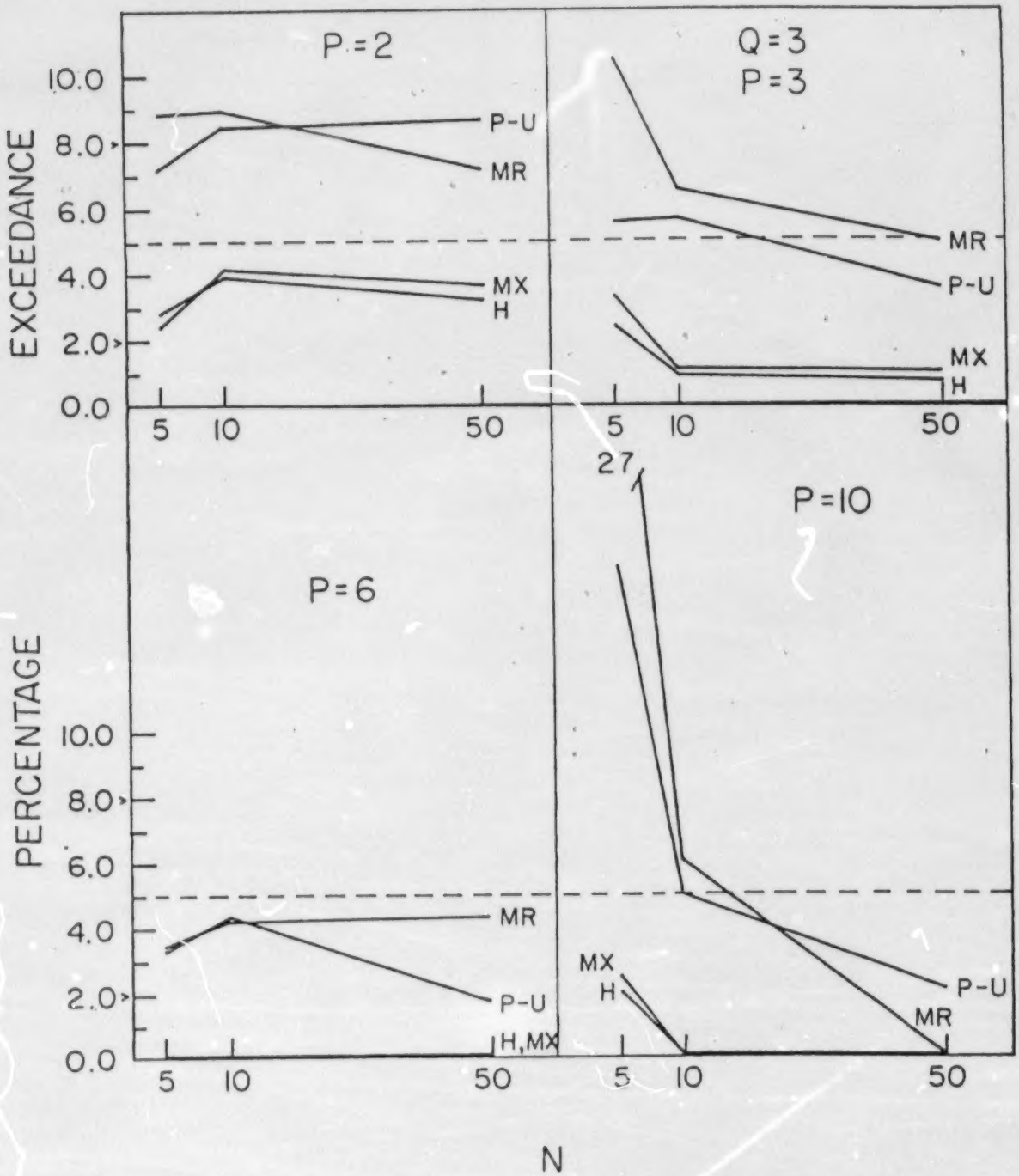

Figure 3(a). Percentage exceedance of theoretical critical values of P-U, H, MX, and MR as tests of the second root in the null case for combinations of p and n given q=3. The nominal significance level of 5% is indicated by the dotted line, while the arrows beside the ordinate denote the boundaries of an approximately two standard error of proportion region around 5%.

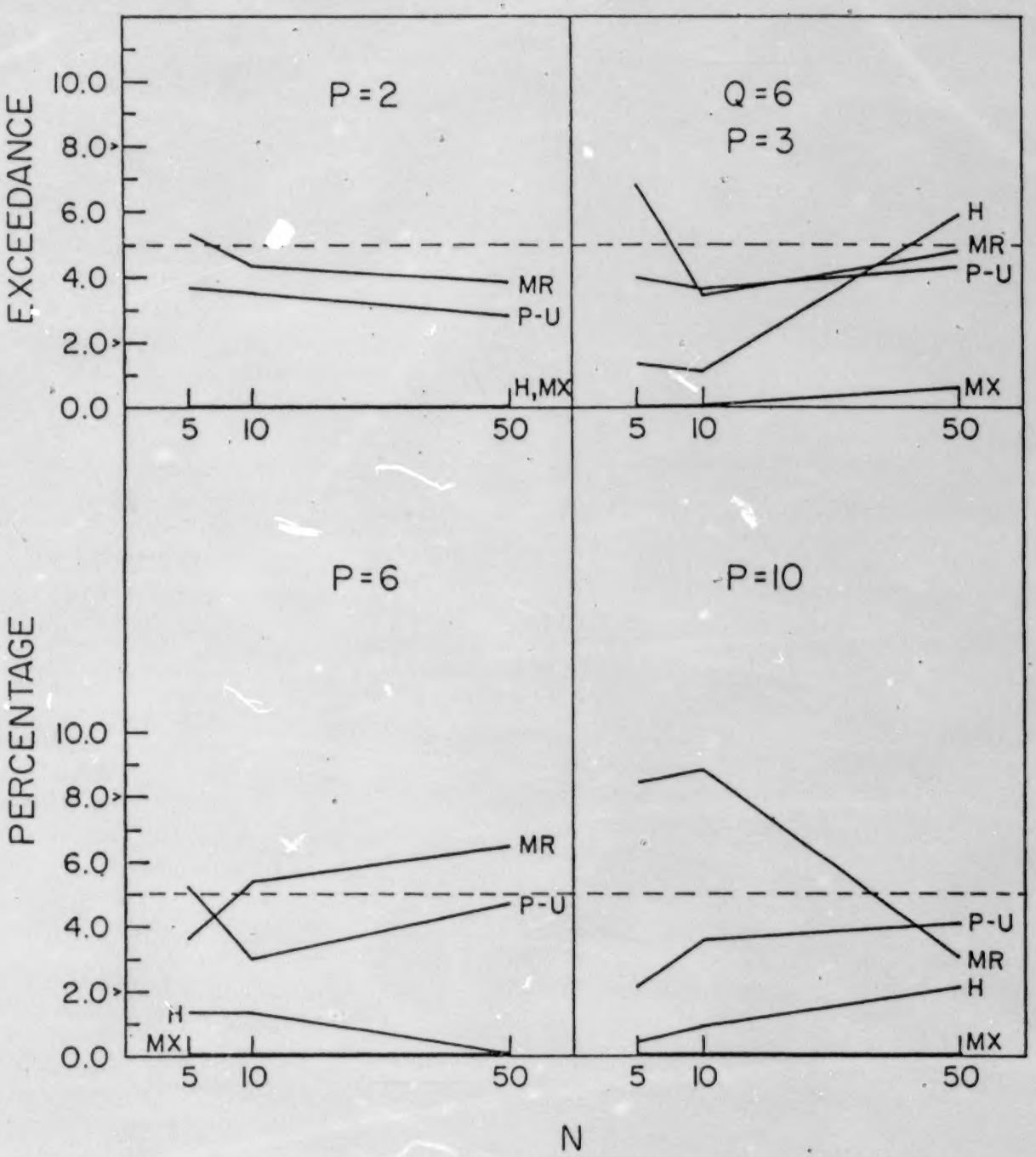

Figure 3(b). Percentage exceedance of theoretical critical values of P-U, H, MX, and MR as tests of the second root in the null case for combinations of p and n given q=6. The nominal significance level of 5% is indicated by the dotted line, while the arrows beside the ordinate denote the boundaries of an approximately two standard error of proportion region around 5%.

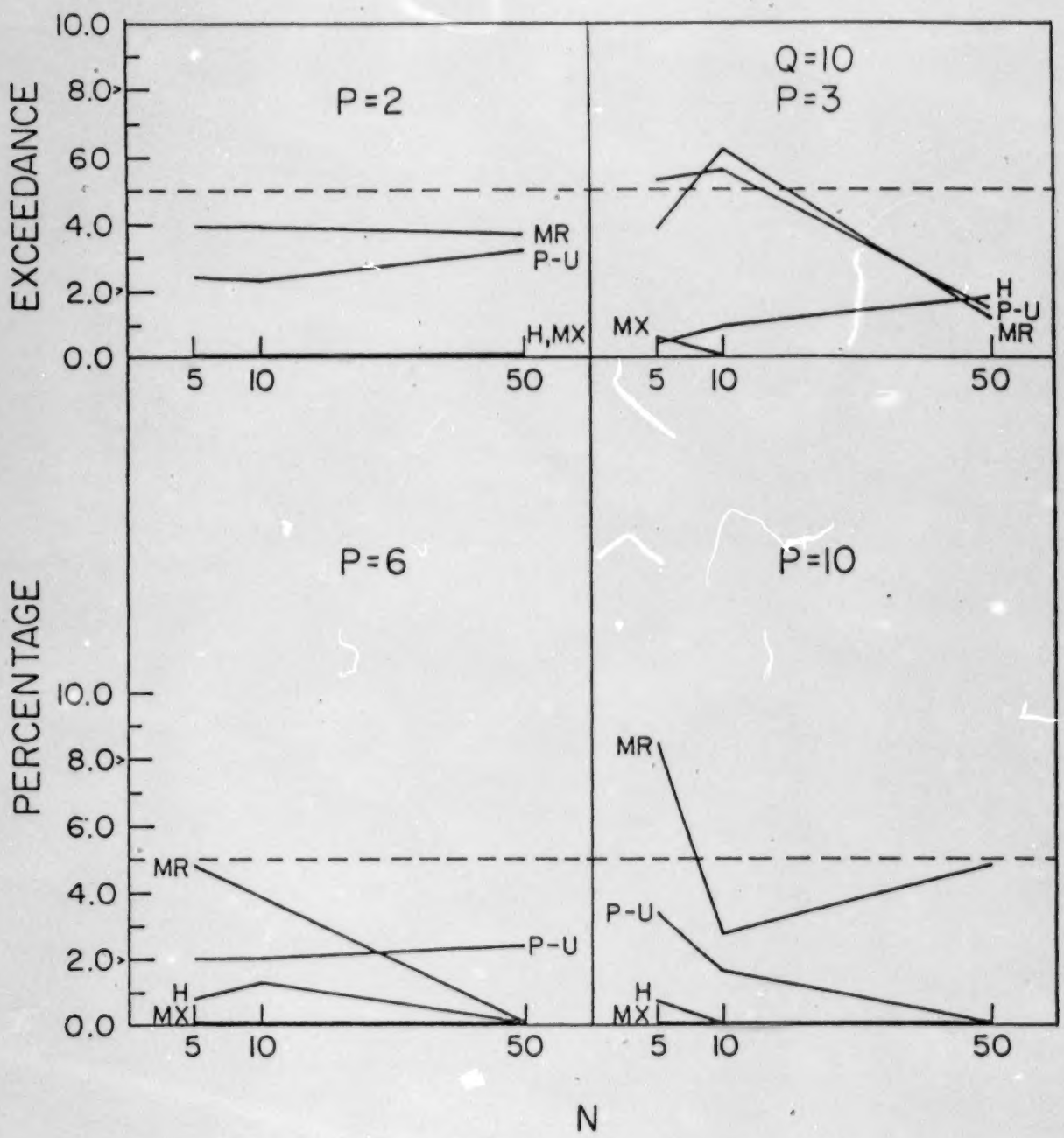

Figure 3(c). Percentage exceedance of theoretical critical values of P-U, H, MX, and MR as tests of the second root in the null case for combinations of p and n given q=10. The nominal significance level of 5% is indicated by the dotted line, while the arrows beside the ordinate denote the boundaries of an approximately two standard error of proportion region around 5%.

of the procedures examined to falsely reject null hypotheses pertaining to the second discriminant function, it was of interest to determine more precisely the behaviour of P-U as a test for roots beyond the first. This interest was generated by Harris' (1976) claim that the procedure would be conservative in its evaluation of roots beyond the first. Accordingly, 200 sets of 1000 replications were generated for condition p=2, q=3, n=10; and 100 sets of 1000 replications were generated for the case of p=3, q=3, n=10. The mean percentage exceedance for the test of the first residual (hence, of λ_2) was found to be 5.7% for the two variate case, and 5.1% for the three variate case. These findings strongly suggest that, for the cases examined at least, P-U is rejecting at or near the nominal rate for both the overall test and the test of the first residual.

Study II. The results reported in this section pertain to the comparative powers of the four procedures to detect nonzero population discriminant functions given various degrees and patterns of multivariate noncentrality. The first cases to be considered are those in which the population noncentrality was concentrated in a single noncentrality root. Figure 4 presents percentage exceedance data for p=6 and combinations of q (3, 6 and 10), n (5 and 10) and the multivariate noncentrality parameter (10, 40 and 90) at a nominal significance level of .05. Inspection of this figure reveals that in general, the tests are performing at approximately the same level. Exceptions to this are the cases in which p=6, where H shows consistently greater sensitivity to an intermediate level of noncentrality (tr(H*) = 40). All of the

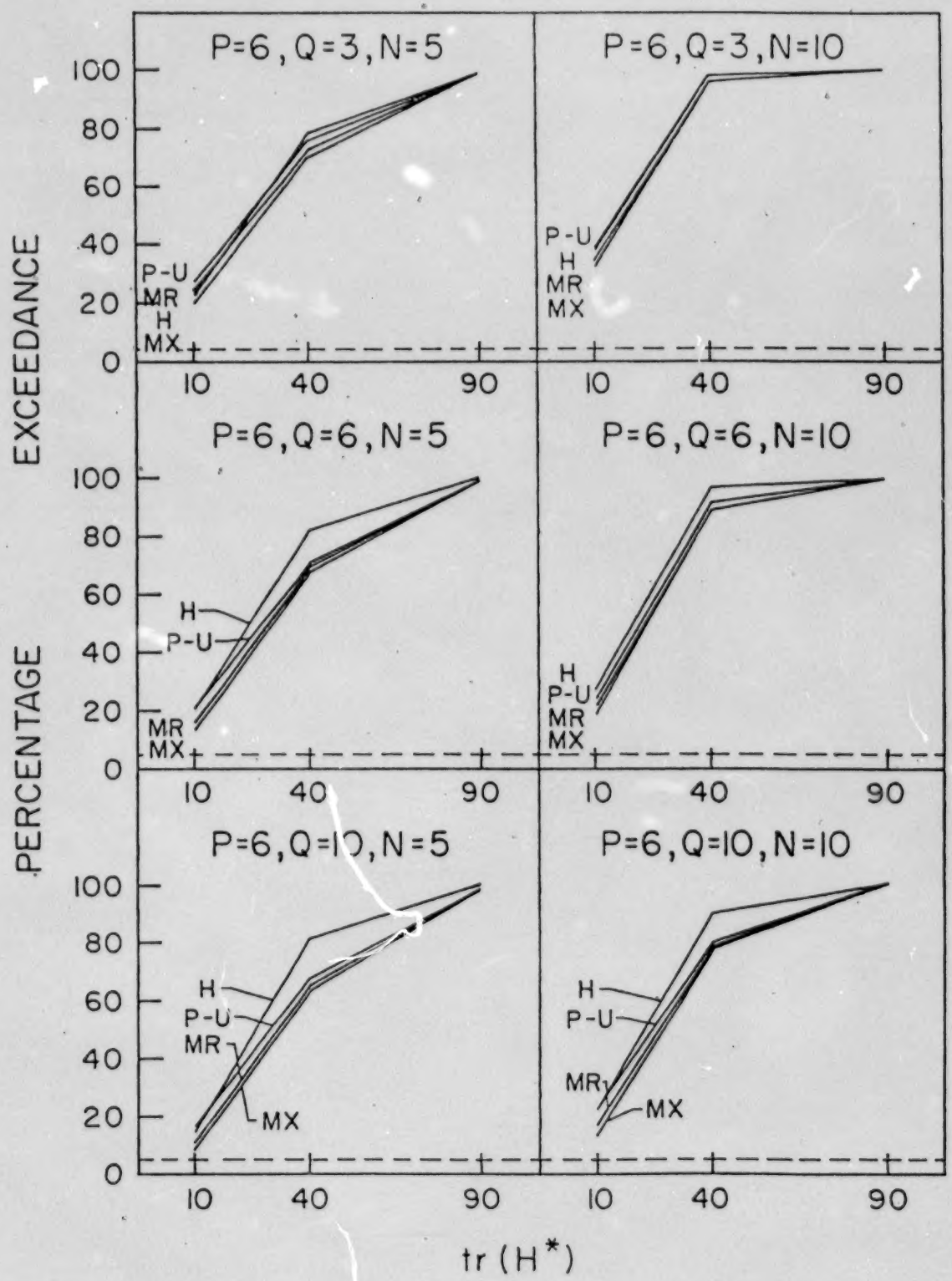

Figure 4. Percentage exceedance of theoretical critical values of P-U, H, MX, and MR as tests of the first root in the concentrated noncentrality case for combinations of p, q, and n as a function of tr(H*). The dotted line denotes the nominal significance level of 5%.

tests rejected H_o in approximately 100% of the samples when the value of tr(H*) simulated was 90, at both nominal significance levels of .05 and .01. The general effects of independent variables may be summarized as follows: percentage exceedance decreased with increases in p or q (hence s), and increased with increases in n, tr(H*) and α . The superiority of H is consistent with earlier findings comparing V and Roy's overall tests among others (e.g., Olson, 1974; Schatzoff, 1966b) which indicated greater power for Roy's test than for V to detect significant discriminant dimensions when all but one of the noncentrality roots are zero.

The results for those cases in which more than one population root was nonzero showed a slight realignment of the four procedures from the concentrated noncentrality pattern. Figures 5(a) to 5(e) present percentage exceedances for selected combinations of p, q and n at a nominal significance level of .05. Note that in cases where two nonzero parameter roots were simulated, only two patterns are presented, equal and linear, since the linear and geometric patterns are identical for two roots. The consistency of the results presented is representative of all of the conditions simulated at both nominal significance levels. The data for each successive root are expressed as a percentage of the number of samples in which the test of the preceding root was significant. For example, if, for a given procedure, the test of λ_1 is significant in 500 of the 1000 replications, then the percentage for the test of λ_2 is taken as a proportion of 500, and not of 1000. Inspection of figure 5 will reveal the following consistent ordering of the four procedures: for all but the last nonzero population root in any set, P-U is the most

powerful procedure followed by MR, H and MX, with the latter three tests generally closer to each other than they are to P-U. For the test of the last nonzero population root, the alignment changes somewhat, with MR slightly superior to P-U, and both of these tests rejecting more frequently than either H or MX. MX is uniformly the least powerful procedure, particularly for tests of roots beyond the first. In terms of the effects of the independent variables individually, the following general conclusions seem warranted: power is a decreasing function of p and q, and increases with larger values of n, tr(H*) and the nominal significance level. The size pattern of the roots was found to be a less potent factor than anticipated, with little difference evident between the linear and geometric patterns, and slightly higher rejection rates for roots tested later in the equal pattern, than in either of the other two. The relatively small size of this last difference underlines the insensitivity of all of the procedures when faced with equal sized population roots. In all cases, the tests were more sensitive to the first root than to the second, more sensitive to the second root than to the third, and so on, despite the equal size of the parameter roots.

In evaluating the performance of the procedures under investigation in the noncentral case, one might be concerned with the comparative frequency with which the procedures overestimate the true dimensionality of a set of sample data. Stated another way, one might ask: what is the comparative likelihood of falsely rejecting $H_o : \lambda_k \neq 0$ after correctly rejecting $H_o : \lambda_{k-1} = 0$, for the four procedures? Data bearing on this question were examined in the present study for cases in which

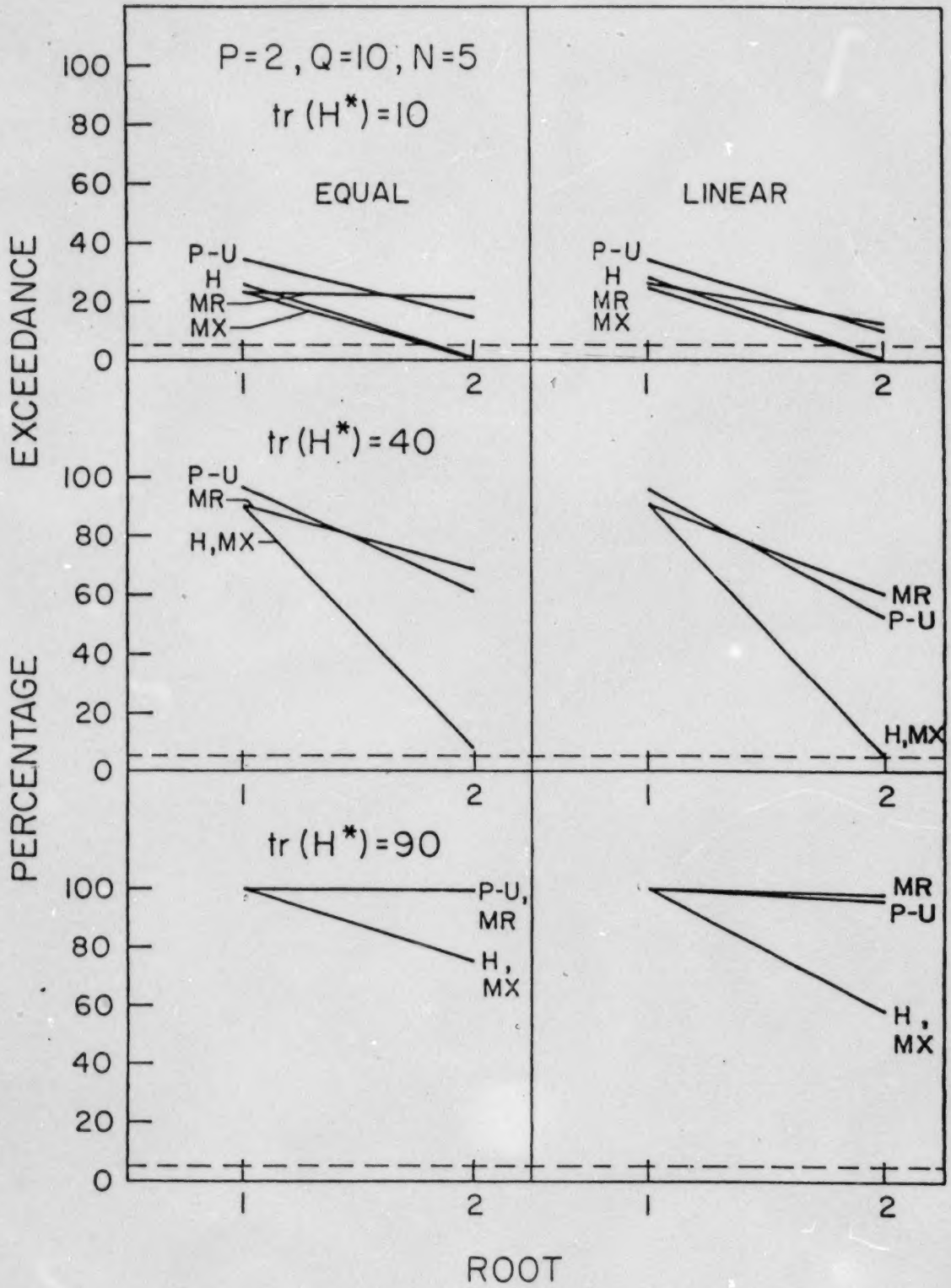

Figure 5(a). Percentage exceedance of the procedures as tests of two non-zero population roots for combinations of tr(H*) and pattern in the case of p=2, q=10, n=5. The dotted line denotes the nominal significance level of 5%.

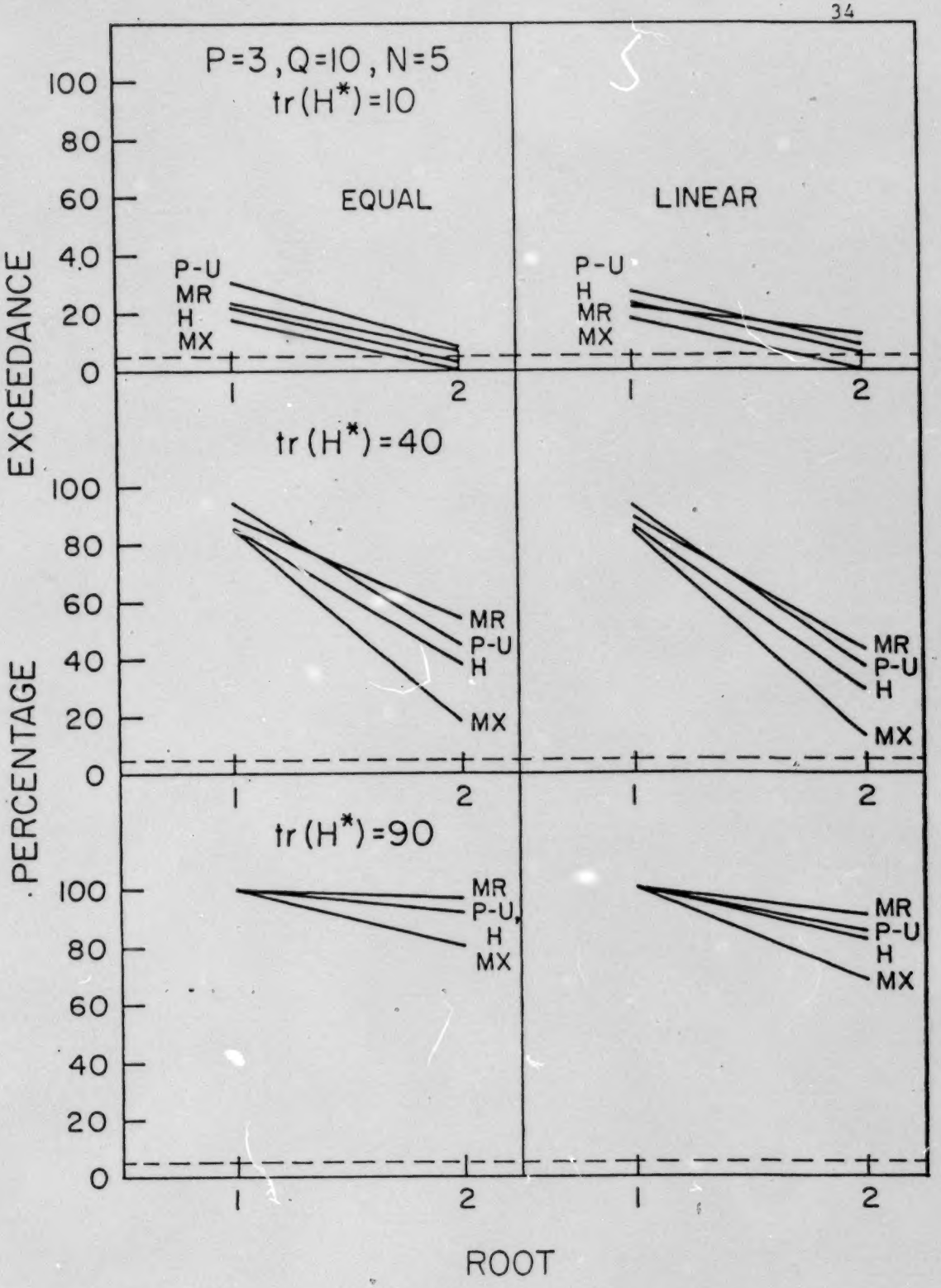

Figure 5(b). Percentage exceedance of the procedures as tests of two non-zero population roots for combinations of tr(H*) and pattern in the case of p=3, q=10, n=5. The dotted line denotes the nominal significance level of 5%.

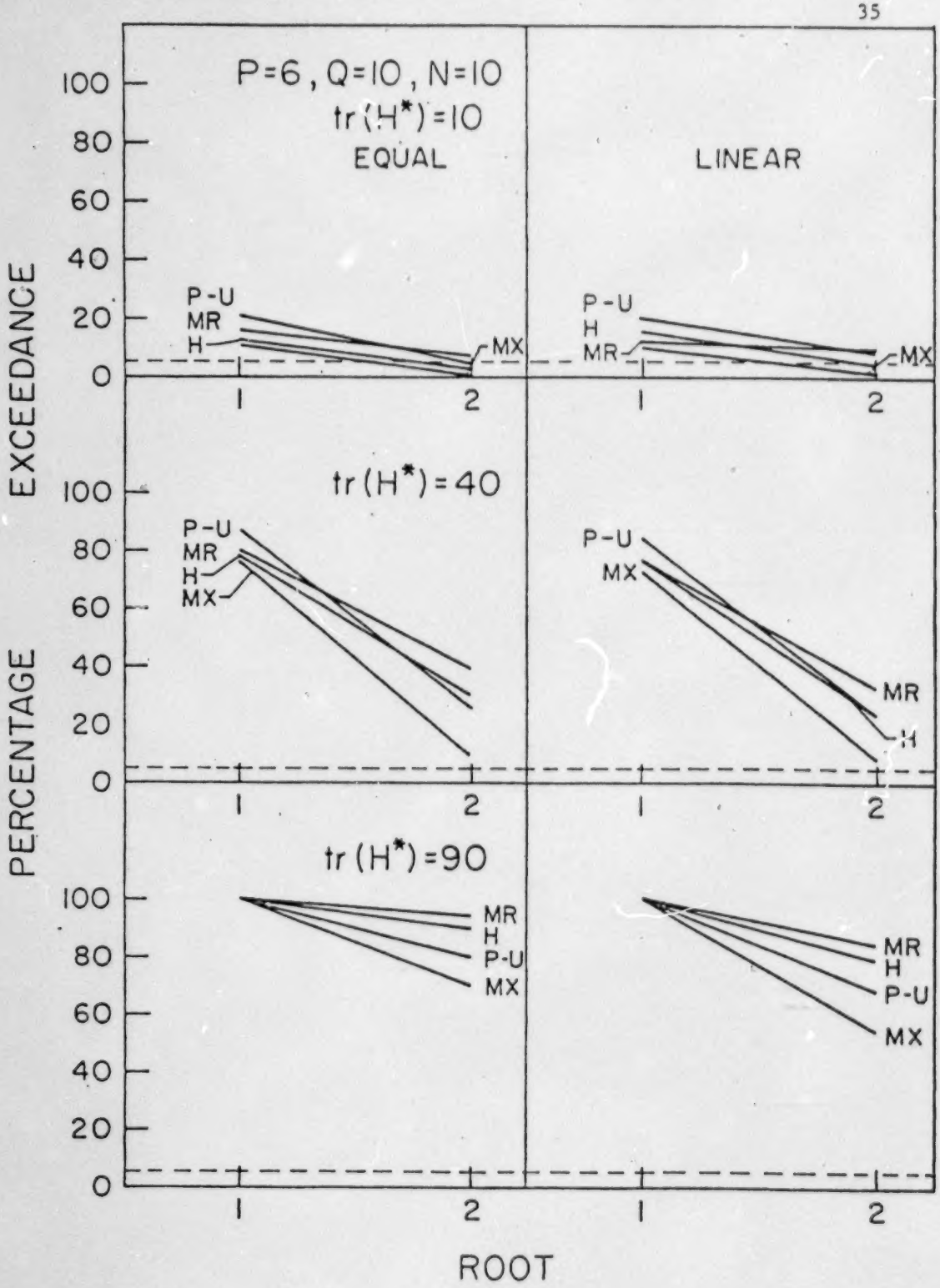

Figure 5(c). Percentage exceedance of the procedures as tests of two non-zero population roots for combinations of tr(H*) and pattern in the case of p=6, q=10, n=10. The dotted line denotes the nominal significance level of 5%.

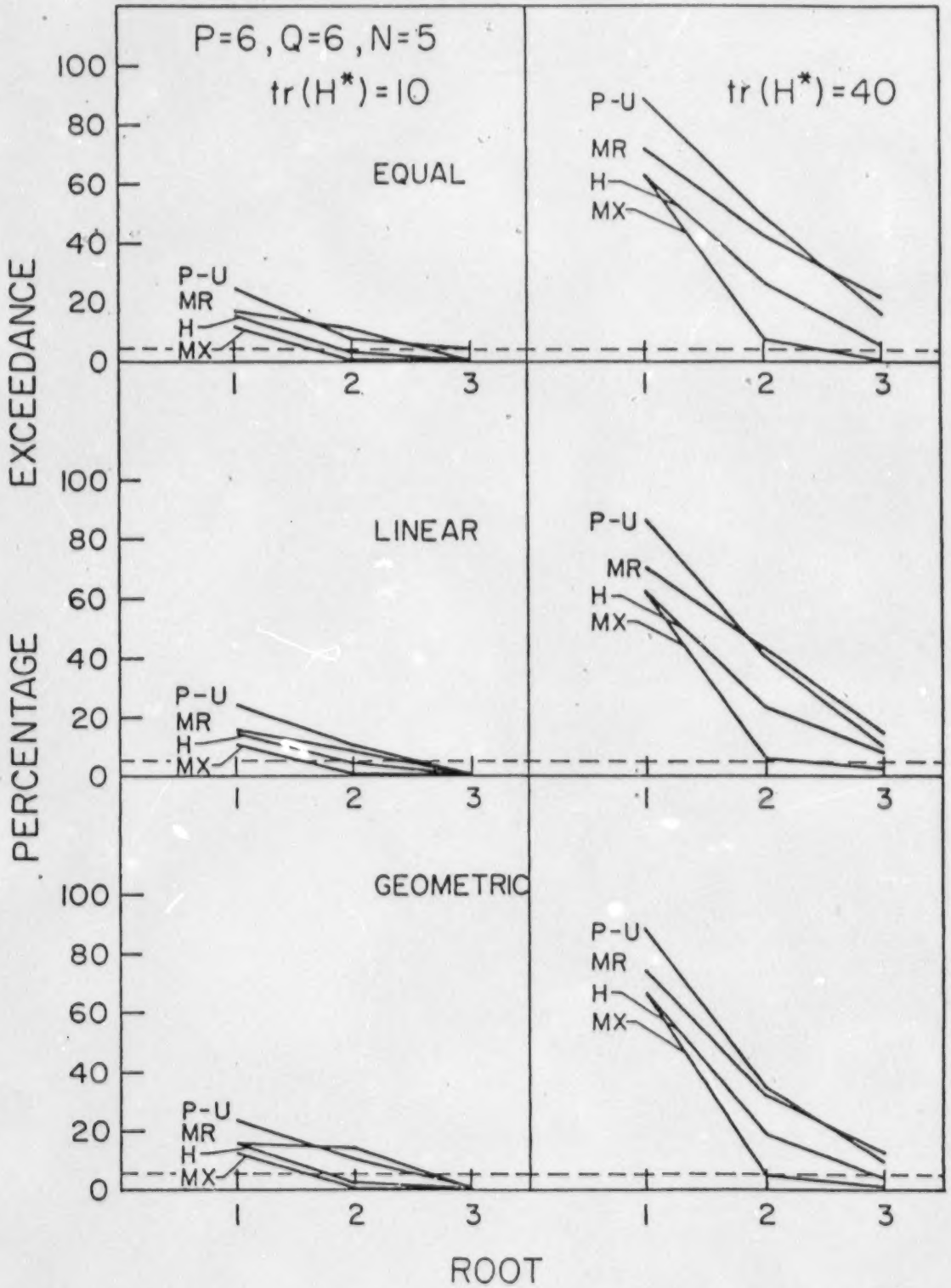

Figure 5(d). Percentage exceedances of the procedures as tests of three nonzero population roots for combinations of tr(H*) and pattern in the case of p=6,q=6,n=5. The dotted line denotes the nominal significance level of 5%.

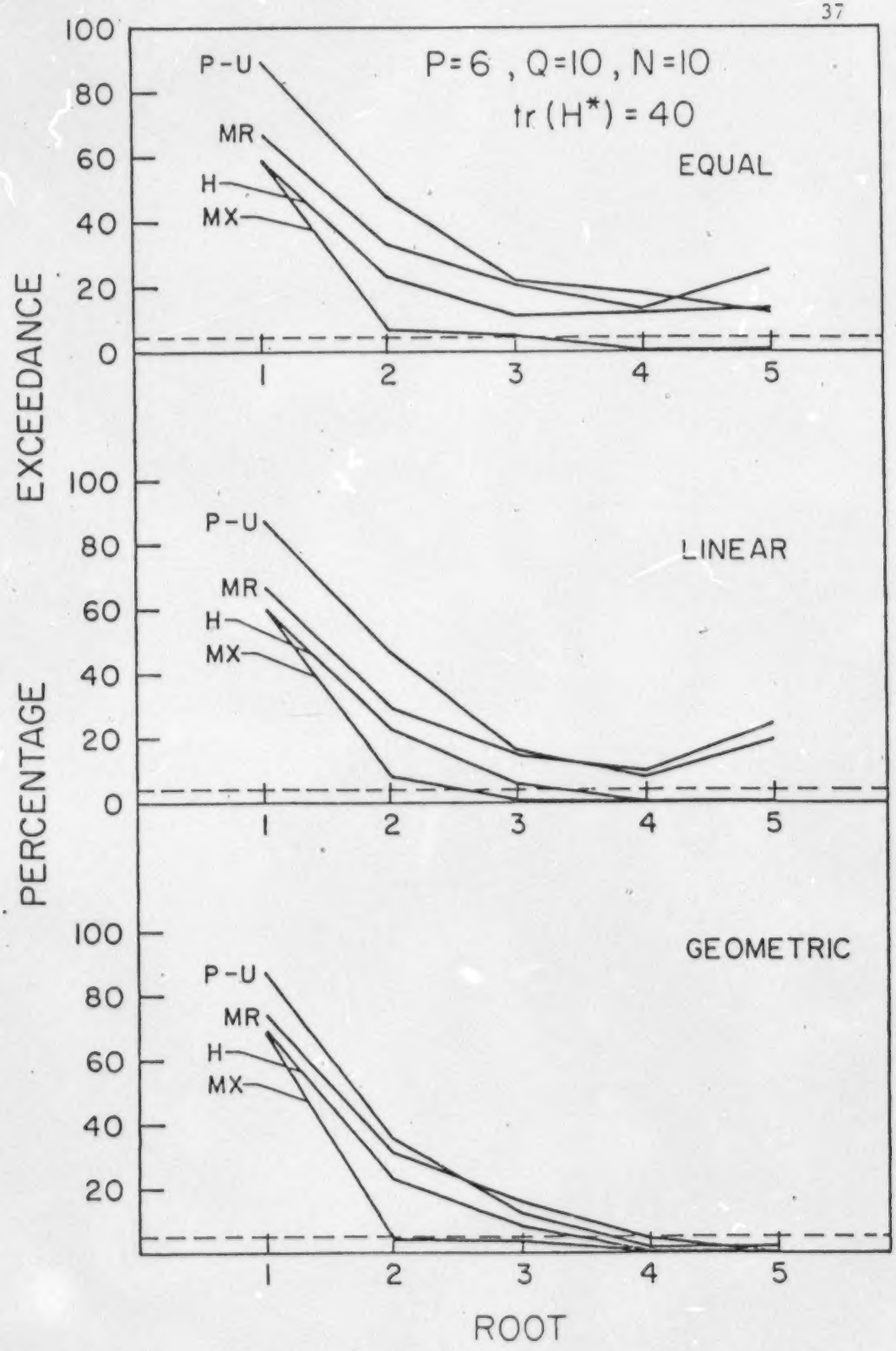

Figure 5(e). Percentage exceedance of the procedures as tests of five nonzero population roots for various patterns in the case p 6,q 10,n 10 and tr(H*) 40. The dotted line denotes the nominal significance level.

dimensionalities of two and three were simulated. The results were similar to those reported for the tests of λ_2 in study I, with P-U and MR rejecting at or near the nominal rate, and, MX and H rejecting at a level well below the nominal rate. In general, the likelihood of such false rejections was a decreasing function of p and q.

Computational Checks. In any simulation study it is imperative to ensure that the data generation procedure is working properly, and yielding reasonable results. In the present study this was accomplished by comparing the rejection rates of Bartlett's V and Roy's overall MANOVA tests to the anticipated theoretical rates in the null case. The direct comparison is feasible in the present study since no violation of distributional assumptions was incorporated into the simulation procedure. With few exceptions, the empirical rejection rates were found to correspond closely to the theoretical rates. The exceptions to this general finding frequently involved Bartlett's V in cases where n was small. Over all combinations of p, q and n, the mean rejection rate at $\alpha = .05$ was 5.34% for V, and 5.1% for Roy's test. Certain comparisons were also possible in the noncentral case. Specifically, Roy (1966) tabulated values of the power function for V for values of tr(H*) - 10, 40 in the concentrated noncentrality situation. Comparison with the results of the present study showed consistent agreement to within approximately two percentage points.

Discussion

The overall results of this study may be summarized as follows: in the null case P-U and H closely approximate the nominal type I error rate, while MR and MX are conservative as tests of the first discriminant function. As tests of the second root (again in the null case) P-U and MR perform at a level more consistent with the nominal rate than do H and MX, which are again conservative. In the noncentral case, P-U generally detects true population discriminant functions as frequently, or more frequently than the other procedures. The only major exception to this is the concentrated noncentrality situation in which H is slightly superior to P-U. The satisfactory performance of P-U and H as tests of the first root in the null case is not surprising, since, as noted above, as tests of λ_1, these procedures are, in fact, Bartlett's V and Roy's criterion, respectively. It is important to note that neither of these procedures directly tests a null hypothesis of the form $\lambda_1 = 0$, but are instead tests of the overall MANOVA null hypothesis, the rejection of which only indirectly implies that at least one population discriminant function is nonzero.

The results of the present study bear directly on Harris' (1976) criticisms of P-U as liberal in its evaluation of λ_1 and conservative in its evaluation of smaller roots. Unlike Harris' (1976) simulation, the procedure employed in the present study is consistent with the null hypothesis tested by P-U, namely that the first k roots are nonzero, while the remaining s-k are zero. Harris' (1976) simulation tested a null hypothesis of the form $\lambda_1 = 0$, which is not the null hypothesis

tested by P-U (Kshirsagar, 1972). Across almost all combinations of p, q and n examined in the present study, P-U rejected correct null hypotheses (the null case) at a level close to the nominal significance level. The most noteworthy exception to this was the case of p=10, q=3, n=5 in which P-U was liberal in its evaluations of both λ_1 and (especially) λ_2. The finding of a 12% rejection rate by P-U (nominal $\alpha = .05$) for the test of the second root in this case (based on 15,000 replications) reflects the instability of this procedure as p approaches df_e (here, 10 and 12, respectively). Only the 27% rejection rate of MR under the same conditions was farther from the nominal rate of 5%. In any event, it seems clear, on the basis of these results, that P-U, correctly applied, is neither liberal in its evaluation of λ_1, nor conservative in its evaluation of λ_2.

The unsatisfactory performance of the procedure proposed by Finn (1974) documents the danger inherent in ad hoc solutions to problems in inferential statistics. Not only did Finn's (1974) procedure lack distributional theory support, but its application in certain situations is clearly illogical due to the degree of freedom problem discussed above. Users of the widely-distributed multivariate analysis program MULTIVARIANCE (Finn, 1972) should be aware that this procedure is included as part of the discriminant analysis output (values of s, m and n are printed for each discriminant function), and should ignore this information when interpreting their results.

To the extent that comparisons are possible, the results of the present study concur with those reported recently by Mendoza et al (1978). In the latter study three variations of Bartlett's test and one of Harris' proposed procedures (H) were compared. In a limited simulation, the null case, as well as several cases of 3 nonzero parameter roots, were examined for p = 6 and q = 7 (see Table 1). The results of the simulation, as well as a logical analysis of Harris' criticisms of P-U led the authors to support the use of sequential procedures over H. A further finding of interest was the tendency of all of the procedures compared to underestimate the number of nonzero parameter roots. In general, the results of the present study extend and confirm the findings of Mendoza et al(1978).

A principal aim of this research project was to attempt to provide guidelines for researchers who wished to determine the number of reliable discriminant functions present in their data. (An example application of the four procedures considered here is provided in Meredith (1978).) The overwhelming recommendation is to use P-U provided that df_e is large relative to p. Across all combinations of p, q and n investigated (except p=10, df_e = 12) P-U was found to be neither unacceptably liberal nor conservative in its rate of false rejections (type I error rate), and to have adequate power against a variety of alternative hypotheses. In addition to these factors P-U has distributional theory support, is widely

available in statistical program packages, and references the chi-square distribution (available in most statistical texts) in a straightforward and simple manner. This last consideration casts H. and MR in an unfavourable light since use of the published tables of the greatest characteristic root distribution (for H) often requires substantial interpolation, and occasionally extrapolation, while MR requires calculation of non-integral degrees of freedom for referencing the chi-square critical values, again with frequent interpolation. It seems reasonable to recommend against use of MX since it lacks distributional theory support, would be tedious to apply in that two sampling distributions would have to be referenced, is conservative in terms of type I error rate, and lacks power.

As noted above, the performance of P-U deteriorated as p approached df_e. Several possible tacks to follow in such situations might be a) use P-U but with a more stringent significance level (say .01 instead of .05) in order to reduce the probability of over-interpreting the number of discriminant functions present in one's data; b) attempt to reduce p by, for example, performing a principal components analysis on the dependent variables prior to the MANOVA; c) increase n; or d) eliminate dependent variables on the basis of theoretical importance or interest.

The recommendation of the present study to employ P-U must be tempered with the following considerations. Importantly, the present study simulated data in which all distributional assumptions were satisfied. Further research will be required to determine whether the procedures considered here are differentially sensitive to assumption

violation. Certainly the results reported by Olson (1973, 1974) indicate that both V and Roy's overall tests are sensitive to violations of the assumption of population variance-covariance matrix homogeneity. It is not clear whether these findings can be extended to P-U and H, but in view of the probable high frequency of violation of this assumption, this question would seem to warrant examination. Another characteristic of the data simulated in this study which is likely atypical of 'real' data is that all groups were equal in size, and the data were complete (i.e., no data were missing). A complete evaluation of the performance of P-U and the other procedures would require simulation of mixed cell sizes and the application of several procedures to compensate for missing data.

References

Barnard, M.M. The secular variation of skull characters in four series of Egyptian skulls. Annals of Eugenics, 1935, 6, 189-195.

Bartlett, M.S. The statistical significance of canonical correlations. Biometrika, 1941, 32, 29-37.

Box, G.E.P. and Muller, M.E. A note on the generation of random normal deviates. Annals of Mathematical Statistics, 1958, 29, 610-611.

Clyde, D.J. Multivariate Analysis of Variance on Large Computers. Miami: Clyde Computing Service, 1969.

Cramer, E.M. and Bock, R.D. Multivariate Analysis. Review of Educational Research, 1966, 36, 604-617.

Dempster, A.P. Elements of Continuous Multivariate Analysis. Reading: Addison-Wesley, 1969.

Dixon, W.J. (ed.). BMDP: Biomedical Computer Programs. Berkeley: University of California Press, 1975.

Finn, J.D. MULTIVARIANCE: Univariate and multivariate analysis of variance, covariance and regression. Ann Harbor: National Educational Resources, 1972.

Finn, J.D. A General Model for Multivariate Analysis. New York: Holt, Rinehart and Winston, 1974.

Fisher, R.A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 7, 179-188.

Gabriel, K.R. A comparison of some methods of simultaneous inference in MANOVA. In P.R. Krishnaiah (Ed.), Multivariate Analysis - II. New York: Academic Press, 1969.

Harris, R.J. A Primer of Multivariate Statistics. New York: Academic Press, 1975.

Harris, R.J. The invalidity of partitioned-U tests in canonical correlation and multivariate analysis of variance. Multivariate Behavioral Research, 1976, 11, 353-365.

Hotelling, H. The generalization of Student's ratio. Annals of Mathematical Statistics, 1931, 2, 360-378.

Hotelling, H. Relations between two sets of variates. Biometrika, 1936, 28, 321-377.

Hotelling, H. A generalized T test and measure of multivariate dispersion. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1951, 23-41.

Huberty, C.J. Discriminant analysis. Review of Educational Research, 1975, 4, 543-598.

Klecka, W.R. Discriminant analysis. In N.H. Nie, C.H. Hull, J.G. Jenkins, K. Steinbrenner and D.H. Bent (Eds.), SPSS: Statistical Package for the Social Sciences (2nd Edition). New York: McGraw-Hill, 1975.

Kshirsagar, A.M. Multivariate Analysis. New York: Dekker, 1972.

Lawley, D.N. A generalization of Fisher's z test. Biometrika, 1938, 30, 180-187 (corrections 467-469).

Marriott, F.H.C. Tests of significance in canonical analysis. Biometrika, 1952, 39, 58-64.

Mendoza, J.L. Markos, V.H. and Gonter, R. A new perspective on sequential testing procedures in canonical analysis: a Monte Carlo evaluation. Multivariate Behavioral Research, 1978, 13, 371-382.

Meredith, C.D. A Monte Carlo investigation of several procedures for testing the significance of individual discriminant functions following multivariate analysis of variance. Doctoral Dissertation, University of Manitoba, 1978.

Miller, R.G. Developments in Multiple Comparisons 1966-1976. Journal of the American Statistical Association, 1977, 72, 779-788.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. and Bent, D.H. (Eds.), SPSS: Statistical Package for the Social Sciences (2nd Edition). New York: McGraw-Hill, 1975.

Olson, C.L. A Monte Carlo investigation of the robustness of multivariate analysis of variance. Doctoral Dissertation, University of Toronto, 1973.

Olson, C.L. Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association, 1974, 69, 894-908.

Olson, C.L. On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 1976, 83, 579-586.

Overall, J.E. and Klett, C.J. Applied Multivariate Analysis. New York: McGraw-Hill, 1972.

Pearson, E.S. and Hartley, H.O. (Eds.). Biometrika Tables for Statisticians. Vol. 2. Cambridge: Cambridge University Press, 1972.

Pillai, K.C.S. Some new test criteria in multivariate analysis. Annals of Mathematical Statistics, 1955, 26, 117-121.

Pillai, K.C.S. Statistical Tables for Tests of Multivariate Hypotheses. Manila: Statistical Center, University of the Philippines, 1960.

Pillai, K.C.S. On the distribution of the largest characteristic root of a matrix in multivariate analysis. Biometrika, 1965, 52, 405-414.

Pillai, K.C.S. Distributions of characteristic roots in multivariate analysis. Part I. Null distributions. The Canadian Journal of Statistics, 1976, 4, 157-184.

Pillai, K.C.S. Distribution of characteristic work in multivariate analysis, Part II. Non-null distributions. The Canadian Journal of Statistics, 1977, 5, 1-62.

Rao, C.R. The ultilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society, Series B, 1948, 10, 159-193.

Rao, C.R. Linear Statistical Inference and its Applications. New York: Wiley, 1965.

Roy, J. Power of the likelihood-ratio test in analysis of dispersion. Multivariate Analysis, 1966, 1, 105-127.

Roy, S.N. The individual sampling distribution of the maximum, the minimum and any intermediate one of the p-statistics on the null-hypothesis. Sankhyà, 1945, 7, 133-158.

Roy, S.N. and Bargmann, R.E. Tests of multiple independence and the associated confidence bounds. Annals of Mathematical Statistics, 1958, 29, 491-503.

Schatzoff, M. Exact distributions of Wilks's likelihood ratio criterion. Biometrika, 1966, 53, 347-358. (a) (Corrections in Biometrika, 1967, 54, 688.)

Schatzoff, M. Sensitivity comparisons among tests of the general linear hypothesis. Journal of the American Statistical Association, 1966, 61, 415-435. (b)

Tatsuoka, M.M. Multivariate Analysis: Techniques for Educational and Psychological Research. New York: Wiley, 1971.

Wilkinson, J.H. The Algebraic Eigenvalue Problem. Oxford: Clarendon Press, 1965.

Wilks, S.S. Certain generalizations in the analysis of variance. Biometrika, 1932, 24, 471-494.

Wishart, J. The generalized product moment distribution in samples from a normal multivariate population. Biometrika, 20A, 32-52.